Semantic Quality through Semantic Definition: Refining the Read Codes through Internal Consistency

Erich B. Schulz M.B. B.S.*, James W. Barrett M.Sc., M.B. B.Chir.,
Colin Price M.Phil., F.R.C.S.
NHS Centre for Coding and Classification, Loughborough, United Kingdom
*Now at National Centre for Classification in Health (Brisbane), QUT, Australia

Checks of internal consistency in controlled medical vocabularies facilitate their development and assist refinement of the underlying terminological model. Two simple checks of consistency between knowledge in the subtype hierarchy and that in semantic definitions of concepts are described. It is proposed that these checks are a helpful adjunct to, but not a replacement for, large-scale involvement of domain experts in construction of controlled vocabularies.

INTRODUCTION

The Read Thesaurus[1,2,3] is a large controlled medical vocabulary of over 150 000 concepts in a directed acyclic graph subtype hierarchy. An important feature of its Version 3 file structure is the facility to semantically define concepts and we have recently documented our experience in the domains of operative procedures[4] and disorders.[5] Semantic definition, that is, the decomposition of complex concepts into their component sub-concepts, enables the vocabulary to be introspective,[6] or capable of self-validation. Additionally, it facilitates translation between vocabularies,[7] user interface construction,[8] generation of alternative hierarchy views,[9] filtering and management of redundancy. Within the Thesaurus, definitions are held as object-attribute-value triples (figure 1), although other representations include semantic networks, conceptual graphs and frames.

This paper describes two simple rules that test the consistency between the semantic definitions and the subtype (or classification) hierarchy. Their application and limitations are also discussed, particularly reasons why the identified errors cannot be corrected automatically.

BACKGROUND

The Read Thesaurus, in common with the UMLS[10] and SNOMED,[11] has undergone iterative, evolutionary development. Its content is derived from a number of sources including previous Read Code versions, formal classifications[12,13,14] and substantial specialist clinical input. Manual integration of the hierarchies from these differing sources with often incompatible axes has been broadly performed pending development of automated methods for completing the task.

Subsequently, a significant proportion of the semantic definition has been achieved, assisted by lexical matching and inheritance across the subtype hierarchy. SNOMED is currently undergoing a very similar refinement process.[15,16] These approaches are useful for the identification of potential relationships between concepts, but manual review is still essential.[4,16]

Traditional knowledge bases often adopt a 'truth maintenance' strategy where each new fact is tested before integration into the knowledge base. Should the new fact be inconsistent it is rejected. This approach is clearly not compatible with the history and magnitude of the Thesaurus in which tests for internal consistency now play a valuable role in rationalisation.

RULES

A key principle in the organisation of the Thesaurus is the maintenance of parallelism between the classification of the objects and that of their intrinsic values. In order to confirm both completeness and correctness of the subtype hierarchy and the semantic definitions two complementary rules test for internal consistency.

The first rule confirms parallelism between object and value hierarchies (figure 2). It states that each

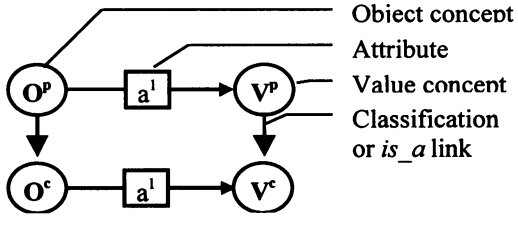
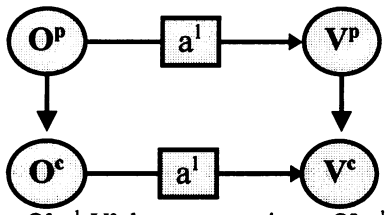


Figure 1 - Key

intrinsic characteristic of a concept must be the same as, or more detailed than, the corresponding characteristic of a hierarchical superordinate.

The second rule attempts to classify concepts based on their semantic definitions (figure 3). In essence, the rule states that a concept with characteristics more detailed than those of another concept must be a subordinate of that concept. Before this rule is applied, the potential superordinate must have been fully defined with respect to a common ancestor.



For every O^c-a^1-V^c there must exist an O^p-a^1-V^p where O^c is an immediate subordinate of O^p and V^c is equal or subordinate to V^p.

Figure 3 – First rule confirming parallelism between the classification hierarchies of object and value concepts. It states that given O^c is a type of O^p, V^c must be a type of V^p.

IMPLEMENTATION

Both rules are implemented in an Oracle™ relational database management system used to maintain the Read Thesaurus. An existing technical quality assurance system composed of a database of approximately five hundred integrity rules provides a scheduling and reporting infrastructure. Most rules in this system are implemented as standard SQL statements. For more complex rules, such as those testing internal consistency, Oracle procedural extensions to SQL (PL/SQL) allow programmatic implementation of appropriate recursive algorithms. An indexing strategy described elsewhere[17] significantly improves execution time.

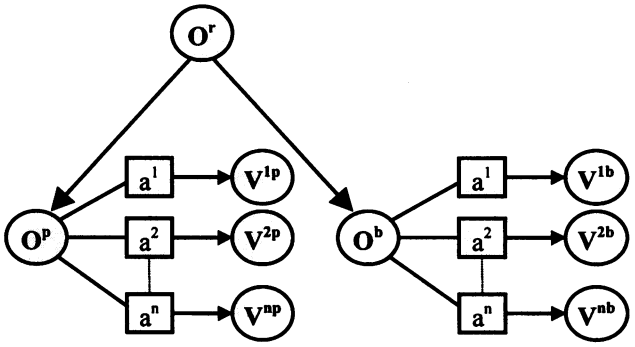
EXPERIENCE

Experience applying these two rules has been mixed. The rules usefully detect errors within the Thesaurus, introduce order and help refine the underlying model. Also, it is generally easier to identify errors of commission rather than errors of omission and the checks of consistency provide a valuable mechanism for detecting the latter. On the other hand, limitations of the underlying representation and the very nature of medical knowledge itself limit the utility of the rules to some extent. Also, semantic definition is labour intensive, even with lexical strategies and inheritance, and requires detailed specialist domain knowledge. There is a danger that resolving the arising inconsistencies becomes a goal in itself, distracting from activities that are more important. Additionally, concept-based tests of internal consistency cannot, by themselves, improve other aspects of terminology quality such as synonym purity[18] or hierarchy induced ambiguity.[3]

Superficially, the two rules appear to simply test for misclassification and for incomplete classification respectively. However, they actually identify several types of inconsistency, including:

- Incomplete or incorrect semantic definition
- Hierarchy misplacement of the object or value concept
- Instability of the still evolving decomposition model[5,19]
- Inconsistent interpretation of lexical cues for disjunction and conjunction.

Figures 4-8 illustrate several patterns of inconsistency that the first rule would detect. These fictitious examples rely on the distinction between the anatomical concepts of *Nail*, *Finger nail* and *Toe nail* and are intended to be illustrative rather than an exhaustive catalogue of scenarios.



O^b must be a subclass of O^p if (1) for every O^p-a^x-V^{xp} there exists an O^b-a^x-V^{xb} such that V^{xb} is equal or subordinate to V^{xp}, (2) O^b and O^p are subordinates of O^r, and (3) the set (a^1-V^{1p} to a^n-V^{np}) captures all differences between O^r and O^p.

Figure 3 – The second rule that detects apparent incomplete classification. It tests whether a classification link is missing between O^b and O^p based on their semantic definitions.

It is the multiple patterns of inconsistency that necessitates human intervention to correct the errors. In the following figures, the location of the error has been highlighted with an 'X'.

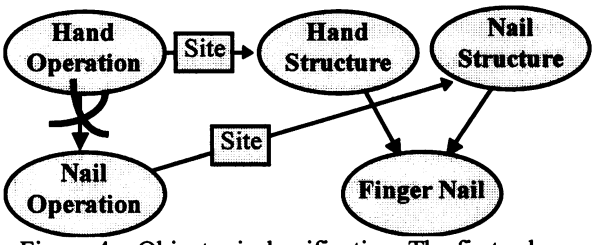


Figure 4 – Object misclassification. The first rule detects that V^c, *Nail Structure*, is not a type of V^p, *Hand Structure*. The true error, however, is that O^c, *Nail Operation*, has been misclassified – because not all nail operations are hand operations.

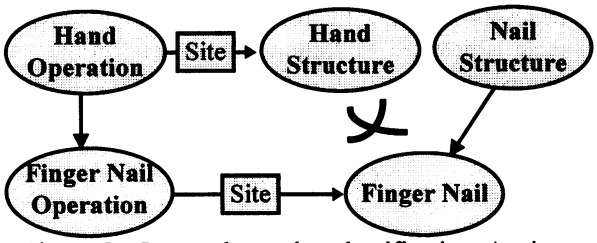


Figure 5 – Incomplete value classification. Again the first rule detects that V^c, *Finger Nail*, is not a type of V^p, *Hand Structure*. In this case the missing classification link needs to be added between *Finger Nail* and *Hand Structure* to correct the error.

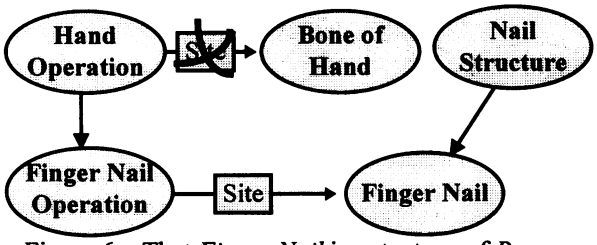


Figure 6 – That *Finger Nail* is not a type of *Bone of Hand* will trigger the first rule, but the real error is misdefinition of the *site* of O^p, *Hand Operation*.

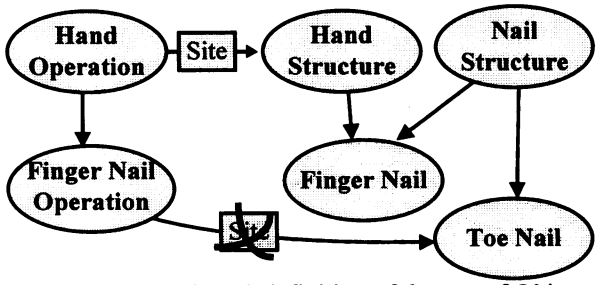


Figure 7 – Here the misdefinition of the *site* of O^c is detected by the first rule because *Toe Nail* is not a type of *Hand Structure*.

Figures 4 and 5 illustrate classification errors in the object and value hierarchies respectively. Figures 6, 7 and 8 show various semantic definition errors.

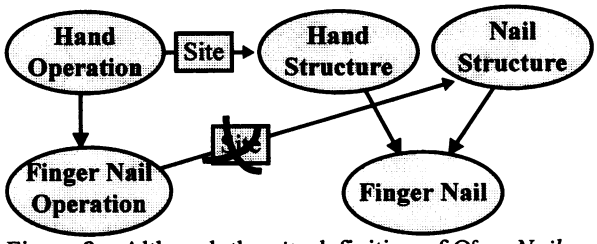


Figure 8 – Although the *site* definition of O^c as *Nail Structure* is not wrong, it is overly general. Again, this is detectable by the first rule because V^c, *Nail Structure*, is not a type of V^p, *Hand Structure*. Refinement of the *site* definition of *Finger Nail Operation* to *Finger Nail* will correct the error.

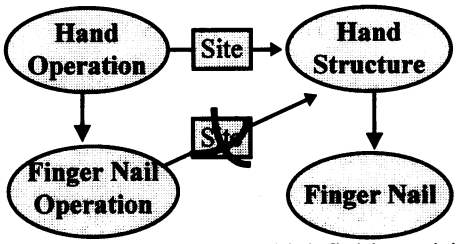


Figure 9 – Over-general child definition arising from unrefined inheritance. This error is not detectable by the first rule, representing a 'blind spot'.

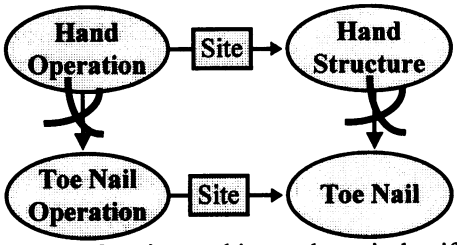


Figure 10 – Consistent object-value misclassification. Because the classification errors in both hierarchies match this situation is also not detectable automatically.

Importantly, some sets of errors are actually internally consistent and are therefore not detectable by confirming object-value parallelism (figures 9-11). Figure 9 illustrates the inappropriate automatic inheritance of definitions and figures 10 and 11 show other scenarios not detectable by the first rule.

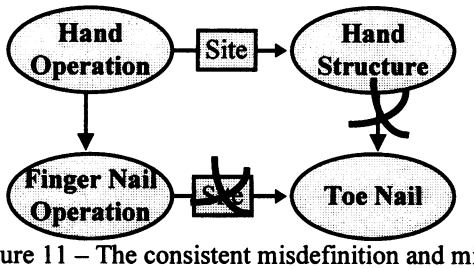


Figure 11 – The consistent misdefinition and misclassification in this figure shows another situation that the first rule cannot detect.

Although the second rule, by definition, detects missing classification links in the Thesaurus (figure 12), overly general semantic definition is also identified. In figure 13, for example, the overly general definition of *Hand Operation* as having a *site* of *Limb structure* rather than *Hand structure* results in *Nail excision* being identified as a 'missing' subordinate of *Hand operation.*

LIMITATIONS

The value of the rules is inherently limited by the expressivity of the underlying data structure. Nesting of semantic definitions, disjunction, conjunction, negation and cardinality are not expressible with the relatively simple object-attribute-value triple formalism of Version 3, so these may not presently be tested.

These limitations could be partially overcome by the adoption of a more expressive notation such as conceptual graphs, GALEN GRAIL,[20] or others. The overheads of adopting a more complex formalism are considerable, however, and require careful thought. Additional costs associated with training Read Code authors, development of editing and quality assurance software, and support of system developers all need justification.

Although not described in this paper, checks for redundancy are also included in the quality assurance system. Detecting duplicate semantic definitions allows identification of incorrect or underspecific semantic definition, or true duplication of concepts. Additionally, redundancy in the hierarchy table is detected by testing that no concept has an immediate parent that is also a non-immediate ancestor.

CONCLUSION

As coding systems become larger, more complex and flexible, quality assurance becomes increas-

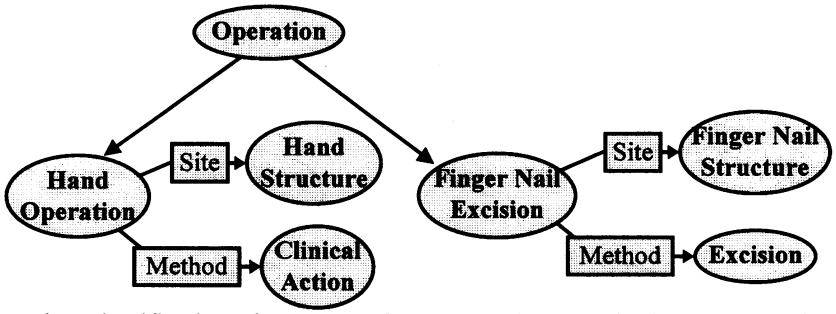


Figure 12 – Incomplete classification of *Finger Nail Excision.* The second rule will detect that O^b, *Finger Nail Excision*, should be classified as a type of O^p because the semantic definition values of *Finger Nail Excision* are classed as types the respective *Hand Operation* values. For clarity, this figure does not show the classification links between the value concepts.

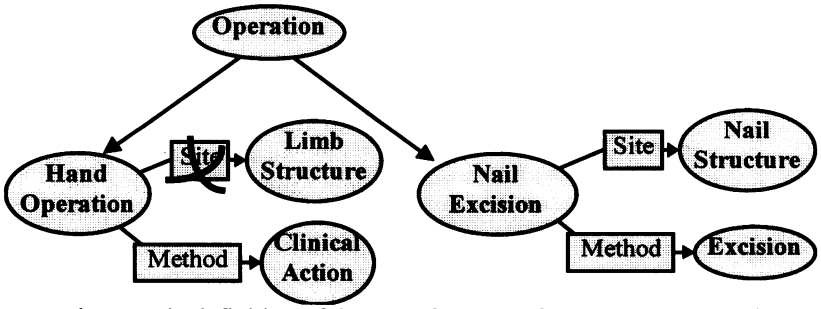


Figure 13 – Over general semantic definition of the *site* of O^p, *Hand Operation*, causes the second rule to incorrectly suggest that O^b, *Nail Excision*, should be classified as a *Hand Operation.*

ingly challenging and must also become an ongoing activity. The ultimate goal of the quality assurance process is to ensure that the products produced by the NHS Centre for Coding and Classification are fit for purpose. This is dependent on completeness, correctness, consistency and conformance to technical specification. Automatic tests of internal consistency are a valuable adjunct to this process. They are not a panacea, however, and detected errors can rarely be corrected automatically. All strategies are dependent on the application of a coherent model and this will undoubtedly prove easier to develop in some domains than others.[5]

Acknowledgments

We thank our reviewers for their helpful comments.

References

1. O'Neil MJ, Payne C, Read JD. Read Codes Version 3: A User Led Terminology. *Meth Inform Med* 1995; **34**: 187-92.
2. Schulz EB, Price C, Brown PJB. Symbolic Anatomic Knowledge Representation in the Read Codes Version 3: Structure and Application. *JAMIA* 1997; **4**: 38-48.
3. Schulz EB, Barrett JW, Brown PJB, Price C. The Read Codes: Evolving a Clinical Vocabulary to Support the Electronic Patient Record. *In Conference Proceedings: Toward an Electronic Health Record Europe. Newton: CAEHR*: 1996: 131-40.
4. Price C, Bentley TE, Brown PJB, Schulz EB, O'Neil MJ. Anatomical Characterisation of Surgical Procedures in the Read Thesaurus. In Cimino JJ (Ed). *Proceedings of the 1996 AMIA Annual Fall Symposium.* Philadelphia: Hanley & Belfus, 1996: 110-4.
5. Brown PJB, O'Neil MJ, Price C. Semantic representation of disorders in Version 3 of the Read Codes. *Meth Inform Med [in press].*
6. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an Introspective, Multipurpose, Controlled Medical Vocabulary. In: Kingsland LC (Ed) *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.* New York: IEEE Computer Society Press, 1989: 513-8.
7. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. *MD Comput* 1990; **7**: 104-9.
8. Kirby J, Rector AL. The PEN&PAD Data Entry System: From prototype to practical system. Cimino JJ (Ed). *Proceedings of the 1996 AMIA Annual Fall Symposium.* Philadelphia: Hanley & Belfus, 1996: 709-713.
9. Cimino JJ, Clayton PD, Hripsack G, Johnson SB. Knowledge based Approaches to the Maintenance of a Large Controlled Medical Terminology. *JAMIA* 1994; **1**: 35-50.
10. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inform Med* 1993; **32**: 281-91.
11. Côté RA, Rothwell DJ, Palotay JL, Becket RS, Brochu L. The systematized nomenclature of human and veterinary medicine - SNOMED International (4 volumes). College of American Pathologists, April 1993.
12. WHO: International Classification of Diseases. 9th Revision. Geneva: WHO, 1975.
13. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Geneva: WHO, 1992.
14. Classification of surgical operations and procedures (4th revision). Office of Population Censuses and Surveys. London: HMSO, 1990.
15. Lipow SS, Fuller LF, Keck KD, Olson NE, Erlbaum MS, Sheretz DD, Nelson SJ. Suggesting Structural Enhancements to SNOMED International. In Cimino JJ (Ed). *Proceedings of the 1996 AMIA Annual Fall Symposium.* Philadelphia: Hanley & Belfus, 1996: 901.
16. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH 1996 Galapagos: Computer-Based Support for Evolution of a Convergent Medical Terminology. In Cimino JJ (Ed). *Proceedings of the 1996 AMIA Annual Fall Symposium.* Philadelphia: Hanley & Belfus, 1996: 269-73.
17. Schulz EB, Smejko N, Price C *et al.* Indexing the Directed Acyclic Graph Hierarchy of the Read Thesaurus. In Cimino JJ (Ed). *Proceedings of the 1996 AMIA Annual Fall Symposium.* Philadelphia: Hanley & Belfus, 1996: 853.
18. Bentley TE, Price C, Brown PJB. Structural and lexical features of successive versions of the Read Codes. In Teasdale S (Ed). *Proceedings of the Annual Conference of the Primary Health Care Specialist Group.* Worcester: PHCSG, 1996; 91-103.
19. Price C, O'Neil MJ, Bentley TE, Brown PJB. Exploring the Ontology of Surgical Procedures in the Read Thesaurus. *Meth Inform Med [in press].*
20. Rector AL, Glowinski AJ, Nowlan WA, Rossi-Mori A. Medical-concept Models and Medical Records: An Approach Based on GALEN and PEN&PAN. *JAMIA* 1995; **2**: 19-35.